Semantic Structure Extraction for Spreadsheet Tables with a Multi-task Learning Architecture

Haoyu Dong¹, Shijie Liu², Zhouyu Fu³, Shi Han¹, Dongmei Zhang¹ ¹Microsoft Research, Beijing, China. ²Beihang University, Beijing, China ³Alibaba Local Service Company (ALSC) Lab, Beijing, China {hadong, shihan, dongmeiz}@microsoft.com, shijie_liu@buaa.edu.cn, zhouyu.fz@alibaba-inc.com

Abstract

Semantic structure extraction for spreadsheets includes detecting table regions, recognizing structural components and classifying cell types. Automatic semantic structure extraction is key to automatic data transformation from various table structures into canonical schema so as to enable data analysis and knowledge discovery. However, they are challenged by the diverse table structures and the spatial-correlated semantics on cell grids. To learn spatial correlations and capture semantic structure extraction. First, we propose a multi-task framework that learns table region, structural components and cell types jointly; second, we leverage the advances of the recent language model to capture semantics in each cell value; third, we build a large human-labeled dataset with broad coverage of table structures. Our evaluation shows that our proposed multi-task framework is highly effective that outperforms the results of training each task separately.

1 Introduction

Spreadsheets are the most popular end-user development tool for data management and analysis. Unlike programming languages or databases, no syntax, data models or even vague standards are enforced for spreadsheets. Figure1(a) shows a real-world spreadsheet. To enable intelligent data analysis and knowledge discovery for the data in range B4:H24, one needs to manually transform the data to a standard form as shown in Figure1(e). It would be highly desirable to develop techniques to extract the semantic structure information for automated spreadsheet data transformation.

Semantic structure extraction entails three chained tasks to: (1) detect tables and locate their respective ranges, shown with red rectangular boxes in Figure1(b); (2) split each table into different components, highlighted with different colors in Figure1(c); (3) categorize individual cells into their corresponding cell types depending on their roles in the transformed table, encoded with different colors in Figure1(d). We also show the transformed data in Figure1(e), where different cell types are highlighted using the same coloring scheme as in Figure1(d).

Learning the semantic structure for spreadsheets is challenging. While table detection is confounded by the diverse multi-table layouts, component recognition is confounded by the various structures of table components, and cell type classification requires semantic-level understanding of cell values. Moreover, the tasks are chained in the sense that latter tasks need to leverage the outcomes of prior tasks. This poses challenges on preventing error propagation, but also provides opportunities for utilizing additional cues from other tasks to improve the current task. For example, header extraction may help table detection since headers need to be inside the table region and vice versa.

In this paper, we present a multi-task learning framework to solve spreadsheet table detection, component recognition, and cell type classification jointly. Our contributions are as follows:

Separ been	Country			mpany 54	85 8,7	IS Tota	Variance		Reportment Com	der a			Company	54 95	M.V. 15	Total	be lance					
		lanary April July	October J	insery 43	15 01	13 156	27	5		denaary		July October	January	43 15	01 10	156	27	1	A	8	С	D
		fost Cest Cest		ibruary 27 arah 30	105	12 144	100	6		Gost	Cust 0	Oest Oest	February March	27 105	12	144	10	1	Departmen	t Country	Month	Cost
Sales	13	23.766 \$ 30.156 \$ 453	\$ 823 A	eil 15		7 22	15	8	Sales B	\$ 23,766	\$ 30,155 \$	453 \$ 823	April	15	7	22	15	2	Sales	US	January	\$ 23.7
	Japan	4.458 3.5.256 3.254	5 522	ry 23 cm 69		14 125	15	2	Jac	at 3 4 658	1 5 250 1	154 5 522		23 0	54	125	15	3	Sales	US	April	\$ 30, 1
				tal 222	129 123 1	64 (29)	94	11					Lteral	222 129	123 154	677	.94	4	Sales	US	July	\$ 45
	others	\$ 050 \$ 23,060 \$ 341		many In	15 5.7	Te Tera	Variance.	12	oth	ire 8.655	\$ 23,060 \$	341 \$ 724	Company	15 15	87 11	Total	by Lanua	5	Sales	Japan	October January	\$ 82
Barbet	15	LANE LAN 1.364		() 41 () 21	45 12	85 33	(4)	14	Barbet B	I AM		104 1 633	July	41 45 21		55 33	(4)	7	Sales	Japan	April	\$ 5.2
Barket				otenber 63		63	21 63	15	Barket B				September	63	12	63	63	8	Sales	Japan	July	\$ 25
	others	\$ 234 \$ 652 \$ 534		taber 22 rember 23	65	97	220	17	oth	ara \$ 234	5 652 5	\$34 \$ 458	Octaber	22 65 23			220	9	Sales	Japan	October	\$ 53
			0	center 17		17	17	19					December	17		17	17	10	Sales	other	January	\$ 65
11	15	1.953 \$ 415 \$ 423	\$ 32	tal 231	110 12	0 353	121	20	17 8	\$ 1,953	\$ 415 \$	423 \$ 32	Tent	231 110	12 0	353	121	11	Sales	other	April	\$ 2.3
	Ohina	8 784 8 205 8 234	1 884					22	ON	5 T54	1 205 1	234 1 884						12	Sales	other	July	\$ 3
	others	5 422 5 205 5 224	6.673					23	oth	6.432	5 205 5	234 5 572						13	Sales	other	October	\$ 7
		144 146 146						10										14	Market	US	January	
								121										15	Market	US	April	\$ 6
(a) T	he origi							121					ble re	egior	ıs bv	red	boxes	15	Market Market	US US	April July	\$ 6 \$ 3
(a) T	'he origi	nal spread						(b) The spr					ble re	egior	ıs by	red	boxes	15	Market	US	April	\$ 6 \$ 3 \$ 6
A	he origi			1 1	K L	A NO	P Q R	(b) The spr	readshee				ble re	egior	is by	red	boxes	15 16 17 18 19	Market Market Market	US US US	April July October	\$ 6 \$ 3 \$ 6 \$ 2
eft Header	he origi			1 1	K L	2 N O	PQR	121	readshee				ble re	egior	is by	red		15 16 17 18 19 20	Market Market Market Market Market	US US others others others	April July October January April July	\$ 6 \$ 3 \$ 6 \$ 2 \$ 6 \$ 5
eft Header	he origi			1 1	K L I	A N O	PQR	(b) The spr	readshee				ble re	egior	is by	red		15 16 17 18 19 20 21	Market Market Market Market Market Market	US US others others others others	April July October January April July October	\$ 6 \$ 3 \$ 6 \$ 2 \$ 6 \$ 5 \$ 5 \$ 4
eft Header op Header	'he origi			1 1	K L	A N O	P Q II.	(b) The spr	readshee				ble re	egior	is by	red		15 16 17 18 19 20 21 22	Market Market Market Market Market Market	US US others others others others US	April July October January April July October January	\$ 60 \$ 34 \$ 60 \$ 20 \$ 60 \$ 50 \$ 50 \$ 50 \$ 50 \$ 40 \$ 51 \$ 40 \$ 51 \$ 51 \$ 51 \$ 51 \$ 51 \$ 51 \$ 51 \$ 51
eft Header op Header lue Region	^b c	nal spread	sheet	I J	K L 1		P Q R	(b) The spr	readshee	t annota	ted w	vith ta		, ,	is by	red		15 16 17 18 19 20 21 22 23	Market Market Market Market Market Market IT IT	US US Others Others Others Others US US	April July October January April July October January April	\$ 61 \$ 61 \$ 3 \$ 61 \$ 2 \$ 61 \$ 5 \$ 61 \$ 61 \$ 61 \$ 61 \$ 61 \$ 61 \$ 61 \$ 61
op Header Iue Region	8 C	nal spread	sheet H	I J Corpory January	43 16 8	1 14	188 27	(b) The spr	readshee	t annota	feed w	with ta) (1 M		P Q k	15 16 17 18 19 20 21 22 23 24	Market Market Market Market Market Market	US US others others others US US US	April July October January April July October January April July	\$ 64 \$ 34 \$ 65 \$ 22 \$ 65 \$ 55 \$ 45 \$ 45 \$ 41 \$ 41 \$ 41 \$ 44 \$ 44
eft Header op Header lue Region	6 C	nal spread	sheet	1 J	43 16 8 27 308 30 4	18	188 27 144 (78) 72 38	(b) The spr	readshee	t annota	feed w	vith ta) K			P 12 8	15 16 17 18 19 20 21 22 23	Market Market Market Market Market IT IT IT	US US Others Others Others Others US US	April July October January April July October January April	\$ 60 \$ 34 \$ 60 \$ 20 \$ 60 \$ 50 \$ 50 \$ 50 \$ 50 \$ 40 \$ 51 \$ 40 \$ 51 \$ 51 \$ 51 \$ 51 \$ 51 \$ 51 \$ 51 \$ 51
op Header lue Region	8 C	nal spread	sheet	1 J	43 16 8 27 309	18	188 27	(b) The spr	readshee	t annota	feed w	with ta	egr Januar) K 100017 00 107 40 107 40 107 40 107 10 10 10 10 10 10 10 10 10 10	L M	a 14 0 14 12 1 7	F Q 8 Intel Decision 156 27 144 (75) 72 28 15	15 16 17 18 19 20 21 22 23 24 25	Market Market Market Market Market IT IT IT IT	US US others others others US US US US	April July October January April July October January April July October	\$ 6 \$ 3 \$ 6 \$ 2 \$ 6 \$ 5 \$ 4 \$ 1 \$ 4 \$ 4 \$ 4 \$ 3 \$ 7
op Header lue Region	6 C	nal spread	Builder Cost	1 J	43 16 8 27 108 4 30 4 15 7 41 8	7 12 12 12 12 12	168 27 164 1700 72 36 22 15 135 15 72 48	(b) The spr	readshee	t annota c o e Country dama con con con con con con con con	ted w s Bast ry Asrit t Cost 10 1 20.156	with ta	Co Bir Pi Site Pi Site Bird Di Ner) X DDATY BA MY 27 13 13 13 13 13 13 13 13 13 13	L M	8 16 0 18 18 12 7 94 12 12 12 12 12 12 12 12 12 13 14 15 15 15 15 15 15 15 15 15 15	F Q 8 1991 Declares 1991 27 144 (7) 72 39 22 15 125 15 72 49	15 16 17 18 19 20 21 22 23 24 25 26 27 28	Market Market Market Market Market IT IT IT IT IT IT IT	US US others others others US US US US US US US US China	April July October January April July October January April July October January	\$ 6 \$ 3 \$ 6 \$ 2 \$ 6 \$ 3 \$ 4 \$ 1 \$ 4 \$ 1 \$ 4 \$ 4 \$ 3 \$ 7 \$ 7 \$ 2
op Header lue Region	6 C serteevi Deutry Salas 23	nal spread	90,5 bleet 00,5 bleet 00 est 1 8 822	1 J	43 16 8 27 308 4 15 23 8	7 12 12 12 12 12	27 168 27 164 (70) 72 30 22 15 135 15	(b) The spr	readshee	t annota c 0 6 Country 26 26 10 10 10 10 10 10 10 10 10 10	F Basel F Auril Cont S 20156 B 5 2.045	any been) X DDATY BA M Y 10 11 13 13 13 13 13 13 13 13 13	L M	8 16 0 18 18 12 7 94 12 12 12 12 12 12 12 12 12 13 14 15 15 15 15 15 15 15 15 15 15	P Q 8 Intel Deciment 19 27 144 (7)8 72 38 22 15 15	15 16 17 18 20 21 22 23 24 25 26 27 28 27 28 29	Market Market Market Market Market IT IT IT IT IT IT IT	US US others others others US US US US China China China	April July October January April July October January April July October January April July October	\$ 6 \$ 3 \$ 6 \$ 2 \$ 6 \$ 5 \$ 4 \$ 1 \$ 4 \$ 1 \$ 4 \$ 3 \$ 7 \$ 2 \$ 2 \$ 2 \$ 3 \$ 5 \$ 4 \$ 5 \$ 4 \$ 5 \$ 4 \$ 5 \$ 5 \$ 4 \$ 5 \$ 5 \$ 5 \$ 4 \$ 5 \$ 5 \$ 5 \$ 5 \$ 5 \$ 5 \$ 5 \$ 5
op Header lue Region	6 C pertment Country Sales 25 Japan	Resth 5 7 6 Janary Arrii Janit Street 6 Janary Arrii Janit Street 6 3 20 Md 3 20 Md 3 20 Md 3 20 Md 1 20 Md 1 20 Md	90,5 bleet 00,5 bleet 00 est 1 8 822	I J Concore Annary, Revolt Agril Bay Anni Felan Felan Felan	43 16 8 27 108 4 30 4 15 4 23 8 49 7 222 129 12 10 85 5	1 18 13 7 94 23 2 154	105 27 144 (70) 72 20 22 15 155 15 72 48 169 94	(b) The spr	readshee	t annota c 0 6 Country 26 121 2000	F Basel F Auril Cont S 20156 B 5 2.045	any acces	Berry Berry Berry Berry David States 12 North 12 North 12 North 12 North 12 North 12 North 12 North 12 North 12 North 12 North 13 North 14) X DDATY BA M Y 10 11 13 13 13 13 13 13 13 13 13	L M E U 15 B 16 17 17 12 12 12 12 12 12 12 12 12	a x a a a a a a a a a a a a a a a a a a	P Q 8 Intel Decision 194 27 144 1736 72 29 125 15 125 15 125 15 126 19 127 19 128 19 129 19 19 19 19 19 19 19 19 19 19 19 19 19 1	15 16 17 17 20 21 22 23 24 25 26 27 28 29 29 30	Market Market Market Market Market IT IT IT IT IT IT IT IT	US US US others others others US US US US China China China other	April July October January April July October January April July October January July October January	\$ 6 \$ 3 \$ 6 \$ 2 \$ 6 \$ 5 \$ 4 \$ 1 \$ 4 \$ 1 \$ 4 \$ 3 \$ 7 \$ 2 \$ 2 \$ 2 \$ 3 \$ 4 \$ 5 \$ 5 \$ 5 \$ 4 \$ 5 \$ 5 \$ 5 \$ 4 \$ 5 \$ 5 \$ 5 \$ 5 \$ 5 \$ 5 \$ 5 \$ 5
eft Header op Header lue Region	6 C pertment Country Sales 25 Japan	Resth 5 7 6 Janary Arrii Janit Street 6 Janary Arrii Janit Street 6 3 20 Md 3 20 Md 3 20 Md 3 20 Md 1 20 Md 1 20 Md	H - - - - - - - - - -	Corpory January Petroary Revels Agent Jana Fata 1 Corpory Jana Fata 1	43 14 8 277 309 4 350 4 4 233 8 4 49 12 129 12 10 85 8 41 45 5 21 8 5	1 18 13 1 7 N4 23 2 154 Y 16	108 27 144 1700 72 38 125 15 135 48 6070 94 50441 Yer ranso 50 60 50 21	(b) The spr	readshee	t annota c 0 6 Country 26 26 10 10 10 10 10 10 10 10 10 10	F F F F F F F F F F F F F F	any been	Constant) (ngany a rr () ngany a rr () ngany a ngany a ng ngany a ngany a ngany a ngany a	L M E Su 15 E 195 195 195 195 195 195 195 195	4 N 0	F Q 8 Intel Decision 158 27 164 178 172 39 123 15 125 15 172 49 164 94	15 16 17 18 19 20 21 22 23 24 25 26 26 26 27 28 29 30 31 1	Market Market Market Market Market IT IT IT IT IT IT IT IT IT	US US others others others others US US US US China China China other other	April July October January April July October January April July October January April July October	\$ 66 \$ 33 \$ 60 \$ 22 \$ 60 \$ 32 \$ 40 \$ 33 \$ 42 \$ 41 \$ 41 \$ 42 \$ 3 \$ 77 \$ 20 \$ 22 \$ 50 \$ 42 \$ 52 \$ 53 \$ 42 \$ 53 \$ 52 \$ 53 \$ 52 \$ 53 \$ 52 \$ 53 \$ 55 \$ 55
eft Header Top Header alue Region	B C Sales 25	Imal spread t r c densrr Krith densrr densrr Krith densrr	H - 00,500rr Cost 1 8 823 1 5 532 5 724	1 1 Concorry Annary February Ref Reg Agenti Reg A	43 14 8 27 108 30 30 4 4 13 8 6 223 8 6 223 729 12 10 85 9 41 45 5 63 5 5	1 18 13 7 23 2 154 2 54	105 27 144 (70) 72 28 22 15 155 15 72 48 609 94 644 The ranks	(b) The spr	readshee	t annota c 0 6 Contry 26 142 28 142 29 142 29 142 29 142 20 14 20	F 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	vith ta 6 H 2 Any 800 6 Any 80) E mpany isa atry 22 b 1 15 20 1 222 mpany isa 1 222 mpany isa 1 222 atr 2 22 1 222 atr 2 22 1 222 atr 2 22 1 222 atr 2 22 1 222 atr 2 22 1 222 atr 2 22 1 222 1 22 1 222 1 22 1 222 1 222 1 222 1 222 1 222 1 222 1 22 1 222 1 222 1 1 1 1	1 M	4 5 0 14 14 17 17 17 14 17 17 14 17 14 17 14 17 17 14 17 17 17 17 17 17 17 17 17 17	P Q 8	15 16 17 19 20 21 22 23 24 25 26 26 27 28 26 27 28 29 30 30 31 32	Market Market Market Market Market Market IT IT IT IT IT IT IT IT IT IT IT IT	US US others others others others others US US US US China China China other other other	April July October January April July October January April July October January April July October January April July	\$ 64 \$ 34 \$ 65 \$ 55 \$ 45 \$ 45 \$ 41 \$ 41 \$ 41 \$ 3 \$ 77 \$ 20 \$ 27 \$ 20 \$ 54 \$ 55 \$ 55 \$ 55 \$ 56 \$ 57 \$ 57 \$ 57 \$ 57 \$ 57 \$ 57 \$ 57 \$ 57
aft Header op Header ilue Region	0 C	t I G desarr Kr11 Joint G desarr Kr11 Joint G 322.00 20.00 20.00 20.00 14.00 31.200 3.20	H - 00,500rr Cost 1 8 823 1 5 532 5 724	1 1 Annary Annary February Bardu Anri 1 Bay Anri 1 Bartal Anri Conseny Anri 1 Bartal Sastaber Destaber Destaber	41 14 8 277 308 4 300 4 4 233 8 4 44 45 5 441 45 5 433 32 -65 323 23 -65	1 18 133 1 7 194 23 2 154 2 154 2 154	Bits 27 104 (700) 77 28 1025 15 1025 15 1025 15 1026 94 1027 28 202 15 1025 15 1026 24 203 21 463 63 47 (33) 33 22	(b) The spr	readshee	t annota	F 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	Anny Book) C mgary ta terr 43 der 22 h 23 h 22 h 24 h 24 h 24 h 25 h 26 h 26	1 M 16 Su 15 S 17 17 17 17 17 17 17 17 17 17	8 5 0 9 N 10 12 7 94 22 154 23 154 23 154 23 154 23 24 25 26 27 27 27 26 27 27 27 27 27 27 27 27 27 27	P Q 8	15 16 17 19 20 21 23 24 25 26 27 28 29 30 31 31 32 33	Market Market Market Market Market IT IT IT IT IT IT IT IT IT	US US others others others others US US US US China China China other other	April July October January April July October January April July October January April July October	\$ 6 \$ 3 \$ 6 \$ 2 \$ 6 \$ 5 \$ 4 \$ 1 \$ 4 \$ 4 \$ 3 \$ 7 \$ 2 \$ 5 \$ 4 \$ 2 \$ 5 \$ 4 \$ 4 \$ 3 \$ 5 \$ 4 \$ 5 \$ 4 \$ 5 \$ 4 \$ 5 \$ 5 \$ 4 \$ 5 \$ 5 \$ 4 \$ 5 \$ 5 \$ 4 \$ 5 \$ 5 \$ 5 \$ 5 \$ 5 \$ 5 \$ 5 \$ 5
eft Header op Header lue Region	B C aartaavi Daurtry Salas 85 Japan 820ar 80rbet 85 s30ar	t r c t r c c domary dom dom dom domary dom dom dom domary dom dom dom dom dom dom dom dom dom dom dom dom dom dom dom dom dom dom dom dom dom dom dom dom dom dom dom dom dom dom	H - Gotaler Cost - 1 800 - 1 800 - 1 800 - 5 400 - 5 400	1 J Corpory January Renth Meril New Jane Telas Corpory May May May May May May May Ma	41 14 8 277 109 9 30 4 4 23 8 4 41 45 9 21 8 9 33 65 17	1 18 133 7 7 194 23 2 154 2 154 2 154	BB 27 164 1701 72 28 1535 15 1525 15 1525 15 164 190 172 48 1829 94 100 21 60 60 57 1330 23 21 17 17	(b) The spr	readshee	t annotal c 0 6 Contry Cont	r r r r r r r r r r r r r r	with ta 6 H any bits See Out Cont Cont 1 400 1 42 540 1 204 540 540 1 500 6.41) 6 10 10 10 10 10 10 10 10 10 10	L M 15 40 15 15 175 47 175 47 175 47 175 122 15 12 15 1	a <u>x</u> 18 12 17 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7	P Q 8	15 16 17 19 20 21 22 23 24 25 26 26 27 28 26 27 28 29 30 30 31 32	Market Market Market Market Market Market IT IT IT IT IT IT IT IT IT IT IT IT	US US others others others others others US US US US US China China China other other other	April July October January April July October January April July October January April July October January April July	\$ 64 \$ 34 \$ 65 \$ 55 \$ 45 \$ 45 \$ 41 \$ 41 \$ 41 \$ 3 \$ 77 \$ 20 \$ 27 \$ 20 \$ 54 \$ 55 \$ 55 \$ 55 \$ 56 \$ 57 \$ 57 \$ 57 \$ 57 \$ 57 \$ 57 \$ 57 \$ 57
aft Header op Header ilue Region	6 C aarteent Courtry Sales 85 28pan	Rest Rest 1 7 6 2 7 7 6 2 0 7 6 2 0 7 6 2 0 7 6 2 0 1 1 1 4 5 3 6 5 0 1 3 6 1 6 1 3 6 5 6 5 5 6 6 5 6 5 6 6 5 6 5 5 6 6 5 6 5	sheet э Элізнег болі в 823 в 724 в 533 в 456 в 537	1 J Corpory January Renth Meril New Jane Telas Corpory May May May May May May May Ma	41 14 8 277 308 4 300 4 4 233 8 4 44 45 5 441 45 5 433 32 -65 323 23 -65	1 18 133 7 7 194 23 2 154 2 154 2 154	BB 27 164 1701 72 28 1535 15 1525 15 1525 15 122 48 1230 21 60 60 233 21 60 63 97 1330 23 22 17 17	(b) The spr	readshee	t annota c 0 6 Courty American Courty	r r r r r r r r r r r r r r	with ta 6 H a H a A a A b A a A b A a A b B Cont c B Cont Cont a A B Cont Cont 1 B Cont Cont Cont Cont 1 B Cont Cont <thcont< th=""> <thcont< th=""> <thcont< th=""></thcont<></thcont<></thcont<>) 6 10 10 10 10 10 10 10 10 10 10	L M 15 40 15 15 175 47 175 47 175 47 175 122 15 12 15 1	8 5 0 9 N 10 12 7 94 22 154 23 154 23 154 23 154 23 24 25 26 27 27 27 26 27 27 27 27 27 27 27 27 27 27	P Q 8	15 16 17 18 19 20 21 22 22 24 24 24 24 24 26 26 26 26 26 29 30 31 32 33 34	Market Market Market Market Market IT IT IT IT IT IT IT IT IT IT IT IT IT	US US others others others US US US US China China China other other other	April July October January April July October January April July October January April July October	\$ 6 \$ 3 \$ 6 \$ 5 \$ 4 \$ 1 \$ 4 \$ 1 \$ 4 \$ 3 \$ 7 \$ 2 \$ 2 \$ 2 \$ 5 \$ 4 \$ 5 \$ 4 \$ 5 \$ 4 \$ 5 \$ 7 \$ 7 \$ 2 \$ 5 \$ 6 \$ 5 \$ 5 \$ 4 \$ 5 \$ 5 \$ 5 \$ 5 \$ 5 \$ 5 \$ 5 \$ 5 \$ 5 \$ 5
ft Header op Header lue Region	B C aartaavi Daurtry Salas 85 Japan 820ar 80rbet 85 s30ar	t r c t r c c domary dom dom dom domary dom dom dom domary dom dom dom dom dom dom dom dom dom dom dom dom dom dom dom dom dom dom dom dom dom dom dom dom dom dom dom dom dom dom	* Ortsher Cost 1 882 1 802 1 802 1 802 1 802 1 803 1 803 1 803 1 803 1 803 1 803	1 J Corpory January Renth Meril New Jane Telas Corpory May May May May May May May Ma	41 14 8 277 109 9 30 4 4 23 8 4 41 45 9 21 8 9 33 65 17	1 18 133 7 7 194 23 2 154 2 154 2 154	BB 27 164 1701 72 28 1535 15 1525 15 1525 15 122 48 1230 21 60 60 233 21 60 63 97 1330 23 22 17 17	(b) The spr	readshee	t annota:	ted w F Bort r Acrit 5 5001 5 1 2015 5 1	with ta 6 H any bits See Out Cont Cont 1 400 1 42 540 1 204 540 540 1 500 6.41) 6 10 10 10 10 10 10 10 10 10 10	L M 15 40 15 15 175 47 175 47 175 47 175 122 15 12 15 1	a <u>x</u> 18 12 17 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7	P Q 8	15 16 17 18 19 20 21 22 22 24 24 24 24 24 26 26 26 26 26 29 30 31 32 33 34	Market Market Market Market Market Market IT IT IT IT IT IT IT IT IT IT IT IT	US US others others others US US US US China China China other other other	April July October January April July October January April July October January April July October	\$ 6 \$ 3 \$ 6 \$ 5 \$ 4 \$ 1 \$ 4 \$ 1 \$ 4 \$ 3 \$ 7 \$ 2 \$ 2 \$ 2 \$ 5 \$ 4 \$ 5 \$ 4 \$ 5 \$ 4 \$ 5 \$ 7 \$ 7 \$ 2 \$ 5 \$ 6 \$ 5 \$ 5 \$ 4 \$ 5 \$ 5 \$ 5 \$ 5 \$ 5 \$ 5 \$ 5 \$ 5 \$ 5 \$ 5

Figure 1: An example that visually illustrates the table semantic structure extraction process.

1. We formulate spreadsheet table structure extraction as a coarse-to-fine process including table detection, component recognition, and cell type classification. We also build a large labeled dataset.

2. To capture the rich information in spreadsheet cells for model training, we devise a featurization scheme containing both hand-crafted features and model-based semantic representations.

3. We propose a multi-task framework that can be trained to simultaneously locate table ranges, recognize table components and extract cell types. Our evaluation shows that the proposed multi-task framework is highly effective that outperforms the results of training each task separately.

2 Problem Statement

Table detection is the task of detecting all tables on a sheet and locating their respective ranges.

Semantic table component recognition is the task of recognizing the top headers, the left headers and the value region in a table. As shown in Figure1(b), a top (left) header contains the labels of table columns (rows) and a value region contains table data (i.e., a body of the table).

Cell type classification is the task of classifying each cell into a certain type such as value, value name, index, and index name. A **value** is a basic unit in the value region. A **value name** is a summary term that describes values. As shown in Figure1(a), "Cost" at E6 is a value name to describe the values in E8:H24. After the data extraction, as shown in Figure1(e), "Cost" at D1 is the label of Column D. An **index** refers to individual values that can be used for indexing data records. In Figure1(a), "January" - "October" at E5:H5 are indexes of columns E - H respectively. A group of indexes is used to breakdown the dataset into subsets. After data transformation, it will form a single data field as Column C shows in Figure1(e). An **index name** is a summary term that describes the indexes. In the previous example, "Month" is the index name of indexes "January" - "October". After data transformation, the "Month" in Figure1(a) corresponds to the column label at C1 in Figure1(e).

3 Method

3.1 Datasets

The web-crawled WebSheet dataset contains 4,290,022 sheets with broad coverage of diverse table structures [1]. We propose a human-labeled dataset, SemanticSheet, for semantic table structure extraction. It includes: (1) 22,176 tables with annotated bounding boxes, which are sampled from WebSheet using an active learning method [1]; (2) 9,053 tables with our structural component annotations, which are randomly sampled from the annotated bounding boxes; (3) 3,503 tables with our cell type annotations, which are sampled using heuristics to balance different table structures based on the structural component annotations. To control labeling quality, all labeled tables have to be verified by an experienced human labeler and be revised until the overall agreement achieves 95%.

3.2 Cell Featurization

To capture effective cell features, we leverage both hand-crafted features and learning-based semantic representations. In general, there are four major information sources of a cell, i.e., value string, data



Figure 2: Framework for spreadsheet table detection, component recognition, and cell classification.

format, cell format, and formula. [1] proposed 20 hand-crafted features and proved its high effectiveness in table detection. But this featurization schema lacks high-level semantic features, which are important for table semantic structure extraction, especially the task of cell type classification.

The recently proposed language models in the natural language processing domain enable to learn embeddings of token sequences with remarkable generalizability. This motivates us to leverage the recent advances in language modeling to derive such high-level semantics in the spreadsheet domain. We incorporate BERT [2] to extract semantic embeddings for cell values. To control the complexity in our model, a point-wise CNN is adopted to reduce the embedding size for each cell value to 32. This CNN model can be jointly trained with the other modules in our framework.

3.3 Multi-task Framework

Figure2 shows our framework, including a shared backbone and three branches for multi-task learning.

Convolutional neural network backbone: Since our featurization method only captures cell-level information, we adopt a Fully Convolutional Neural Network (FCNN) to learn spatial correlations between cells. The FCNN backbone enables our framework to learn shared feature representations, which are able to generalize across multiple different tasks.

Table detection: We adopt and implement the two-stage detection mechanism introduced in [1].

Structural component recognition: This module learns to predict the separation lines between the headers and value region. If there is no top or left header in a table, this module predicts the top and left boundary of the value region. A row-wise average pooling and a softmax function are used to predict the horizontal separation line across rows. And similar with column-wise predictions.

Semantic cell type classification: This module uses a FCNN for cell-level type classification.

Moreover, to leverage the relationships in our three coarse-to-fine tasks, the component recognition branch takes the results of the table detection as an additional feature channel, and the cell type classification branch takes the results of the component recognition as it's additional feature channel.

4 Experiments

4.1 Implementation and Experiment Setup

We take ResNets [3] as the backbone in our framework and exclude all pooling layers to avoid losing cell-level precision. The CNNs for structural component recognition, and cell type classification consist of five convolutional layers. The whole dataset is randomly divided into 80% training set and 20% test set. The loss functions for different tasks are added together for joint training.

4.2 Evaluation Metrics

For table detection, we adopt the Error-of-Boundary (EoB) metric [1] to measure how precisely the detection result is aligned to the ground truth bounding box. For structural component recognition, we calculated the accuracy for top and left header separation lines. For cell type classification, we report the average F1 for the index, index name, and value name predictions.

4.3 Multi-task Framework Evaluation

For the comparison study, we adapt Mask RCNN [4], the state-of-the-art multi-task method for image object detection and segmentation. To evaluate the effectiveness of multi-task joint training, we conduct a comparison between single-task and multi-task training of our proposed method.

Table 1: Co	omparison	results of	spreadsheet	table	semantic stru	acture extraction.

%		Tabel d	etection		Component	recognition	Cell type classification			
	EoB-0		EoB-2		Top header	Left header	Value name	Index	Index name	
Model	Recall	Precision	Recall	Precision	Accuracy	Accuracy	F1	F1	F1	
Mask R-CNN	48.4	40.2	75.5	64.2	56.5	59.1	44.3	65.0	50.2	
Single-task with BERT	81.1	77.4	91.5	87.2	90.4	89.9	70.3	84.6	69.3	
Multi-task with BERT	81.8	78.3	92.1	89.1	91.2	92.3	71.9	85.3	70.8	
Multi-task w/o BERT	81.5	78.3	91.8	89.0	90.1	90.8	66.2	82.2	64.4	

For table detection evaluation, as shown in the left part of Table 1, our method achieves improvements over all baselines. Moreover, the multi-task version of our method performs better than its single-task version in both EoB-0 and EoB-2, indicating that other tasks help with table detection. We attribute such improvements to the learning of intrinsic relationships between these tasks by training a joint model. For the component recognition evaluation, compared with single-task training, our multi-task framework also achieves 0.8% and 2.4% accuracy gains. And the right side of Table 1 shows the results of cell type classification, our method still achieves a large margin of improvement over Mask R-CNN. Compared with single-task training, the joint framework improves the value name, index, and index name predictions by 1.6%, 0.7%, and 1.5% respectively.

4.4 Effectiveness of Semantic Features Extracted by Language Models

We design a comparison experiment to evaluate the effectiveness of language models. For the baseline (multi-task w/o BERT), we only use the 20-dimensional hand-crafted features; while for the treatment (multi-task with BERT), we use the full feature set including the semantic features extracted by BERT. The comparison results are shown in Table 1, indicating that by incorporating the semantic features, our proposed model achieves higher (or on-par) accuracy in all these three tasks. Specifically, since cell type classification heavily depends on semantics, there is 6.4% F1 gain for the index name prediction and 5.7% F1 gain for the value name prediction.

5 Related Work

Some research has been done to extract table data from a spreadsheet [1, 5, 6, 7]. But they only focused on one of the subtasks such as spreadsheet table detection and hierarchical header extraction. And the semantics in spreadsheets have largely been overlooked. We first propose a multi-task framework to learn table detection, component recognition, and cell type recognition simultaneously on a large scale and incorporate language models for capturing semantics in spreadsheets.

References

- [1] Haoyu Dong, Shijie Liu, Shi Han, Zhouyu Fu, and Dongmei Zhang. Tablesense: Spreadsheet table detection with convolutional neural networks. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2019.
- [2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.
- [3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [4] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In Computer Vision (ICCV), 2017 IEEE International Conference on, pages 2980–2988. IEEE, 2017.
- [5] Elvis Koci, Maik Thiele, Óscar Romero Moral, and Wolfgang Lehner. A machine learning approach for layout inference in spreadsheets. In *IC3K 2016: Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management: volume 1: KDIR*, pages 77–88. SciTePress, 2016.
- [6] Kerry Shih-Ping Chang and Brad A Myers. Using and exploring hierarchical data in spreadsheets. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pages 2497–2507. ACM, 2016.
- [7] Zhe Chen and Michael Cafarella. Integrating spreadsheet data via accurate and low-effort extraction. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1126–1135. ACM, 2014.